Analysis of the Distribution of Participation in Wikis Using the Gini Coefficient, the Frequency Distribution and the Lorenz Curve

Javier Arroyo[1,⋆], **Antonio Irpino**[2], **Samer Hassan**[1,3]

1. Facultad de Informática. Universidad Complutense de Madrid, Spain
2. Università degli Studi della Campania "Luigi Vanvitelli", Italy
3. Berkman Klein Center at Harvard University, USA
⋆Contact author: javier.arroyo@fdi.ucm.es

Keywords: Distribution of participation, Distributional data analysis, Lorenz curve, Inequality.

The last two decades have seen the emergence of online communities that produce commons, such as free/open source software or wikis. They are called online peer production communities. Participation in these communities behaves in a peculiar way. While the participation distribution typically exhibits strong levels of inequality across its participants, the collective effort is able to produce goods of high quality and considerable size, such as Wikipedia or Linux to cite some popular examples.

These communities usually follow the 1-9-90 rule (Nielsen, 2006), that is, a 1% of the community population are core contributors concentrating the majority of workload; a 9% are occasional contributors, with sporadic contributions; and a 90% are "consumers", which do not directly contribute, but may do it indirectly, e.g. increasing the visibility of the project. More precisely, the distribution of participation in these communities is said to follow a power-law like distribution.

The distribution of work in online peer production communities is a topic that has attracted a lot of research attention (Kittur *et al.* (2007); Ortega *et al.* (2008); Fuster Morell (2010)). However, the distribution of work is typically summarized by means of the Gini coefficient, or by other metrics that focus in one aspect of the distribution, and ignore the rest of them (Serrano *et al.* (2018)). In fact, the case is similar to that of economic inequality (Cowell (2011)).

Symbolic data analysis make possible to directly study distributions, without relying on descriptive summaries that focus in one aspect of the distribution and ignore the rest of them, offering an incomplete view. In order to do so, distributions need to be represented as histograms (i.e. binned representations). In this work, we will make use of this representations to answer the question of how different are the distributions of participation in online communities. Can we see significantly different participation distributions? Do the shapes depend on the community size or age? Does the shape distribution evolve through time for a given community?

More precisely, we will focus in the participation of contributors in wikis (i.e. editors) and we will compare three approaches for representing the participation distribution in wikis: the Gini coefficient, the distribution of edits across editors, and the Lorenz curve. The first approach is non-symbolic as uses a summary measure, more precisely, an inequality measure. While the last two are symbolic, one using the histogram describing the observed frequency distribution, and the other a special kind of histogram-like distribution, the Lorenz curve which is a cumulative distribution that represents the proportion of overall participation (edits) carried out by the bottom x% of the editors. Unlike the observed participation distribution, the Lorenz curve is a dimensionless distribution bounded by the range [0,1], which ignores the differences in terms of distribution location and

spread and focuses in the distribution shape. We will try to shed light on the above questions analyzing the different participation profiles in wikis with the help of clustering analysis for classic and distributional data (Irpino *et al.* (2006)).

References

Cowell, F. (2011). *Measuring Inequality*. Oxford University Press, Oxford, UK.

Fuster Morell, M. (2010). Participation in online creation communities: Ecosystemic participation. In *Conference Proceedings of the JITP 2010: The Politics of Open Source, Vol. I (Amherst, USA)*, pp. 270–295, Scholarworks@UMassAmherst.

Irpino, A., Verde, R., Lechevallier, Y. (2006). Dynamic clustering of histograms using Wasserstein metric. In *COMPSTAT 2006 - Proceedings in Computational Statistics (Rome, Italy)*, pp. 869–876, Physica-Verlag.

Kittur, A., Chi, E.H., Pendleton, B.A., Suh, B., Mytkowicz, T. (2007). Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. In *CHI '07 - Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, USA)*, ACM.

Ortega, F., Gonzalez-Barahona, J.M., Robles, G. (2008). On the Inequality of Contributions to Wikipedia. In *HICSS 2008 - Proceedings of the 41st Annual Hawaii International Conference on System Sciences (Walkoloa, USA)*, pp. 304–304, IEEE.

Nielsen, J. (2006). The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities `https://www.nngroup.com/articles/participation-inequality/`.

Serrano, A., Arroyo, J., Hassan, S. (2018). Participation Inequality in Wikis: A Temporal Analysis Using WikiChron. In *OpenSym '18 - Proceedings of the 14th International Symposium on Open Collaboration (Paris, France)*, pp. 12, ACM.